Predicting Football Match Outcomes

Machine Learning Project - Bar Ilan University

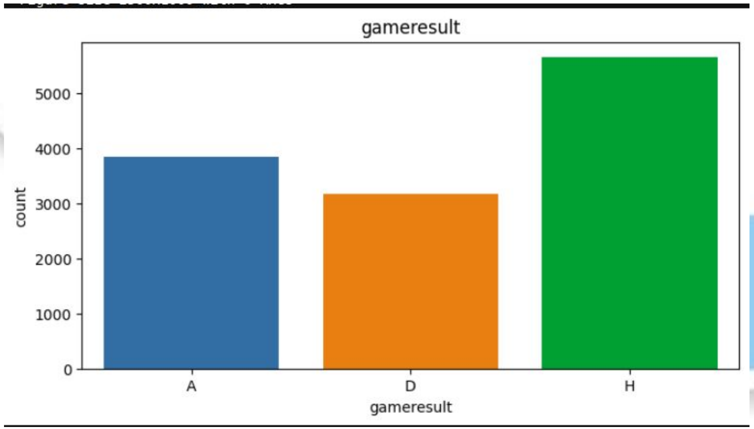
- **Presented by:** Leonardo Romano
- **Course:** Machine Learning (Spring 2025)

Challenges

- Separated datasets
- Selecting the working environment
- Computer crashes
- Project workflow not so smooth

Executive Summary

Objective: Predict match result (Home Win / Draw / Away Win)



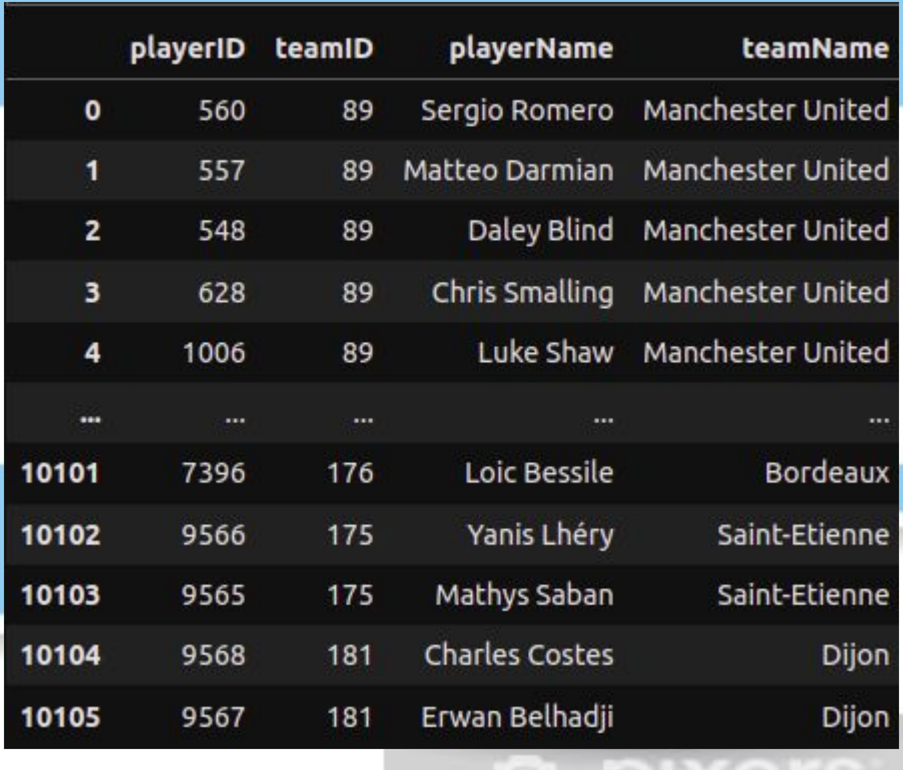
- Dataset from Kaggle: https://www.kaggle.com/datasets/technika148/football-database
- Multi-step pipeline:
    - Data cleaning & feature engineering
    - Exploratory analysis & statistical testing
    - Model training & tuning
- Final model: XGBoost Classifier with >99% accuracy

Procedure

1. Linked PlayerID to TeamID
2. Date parsing & time feature extraction
3. EDA
4. Creating categorical features to handle outliers with low cardinality
5. Run MICE to fill missing values due to outliers removal
6. Feature Engineering and Selection
7. Model selection and final Model used XGboost

Procedure

1. Linked PlayerID to TeamID



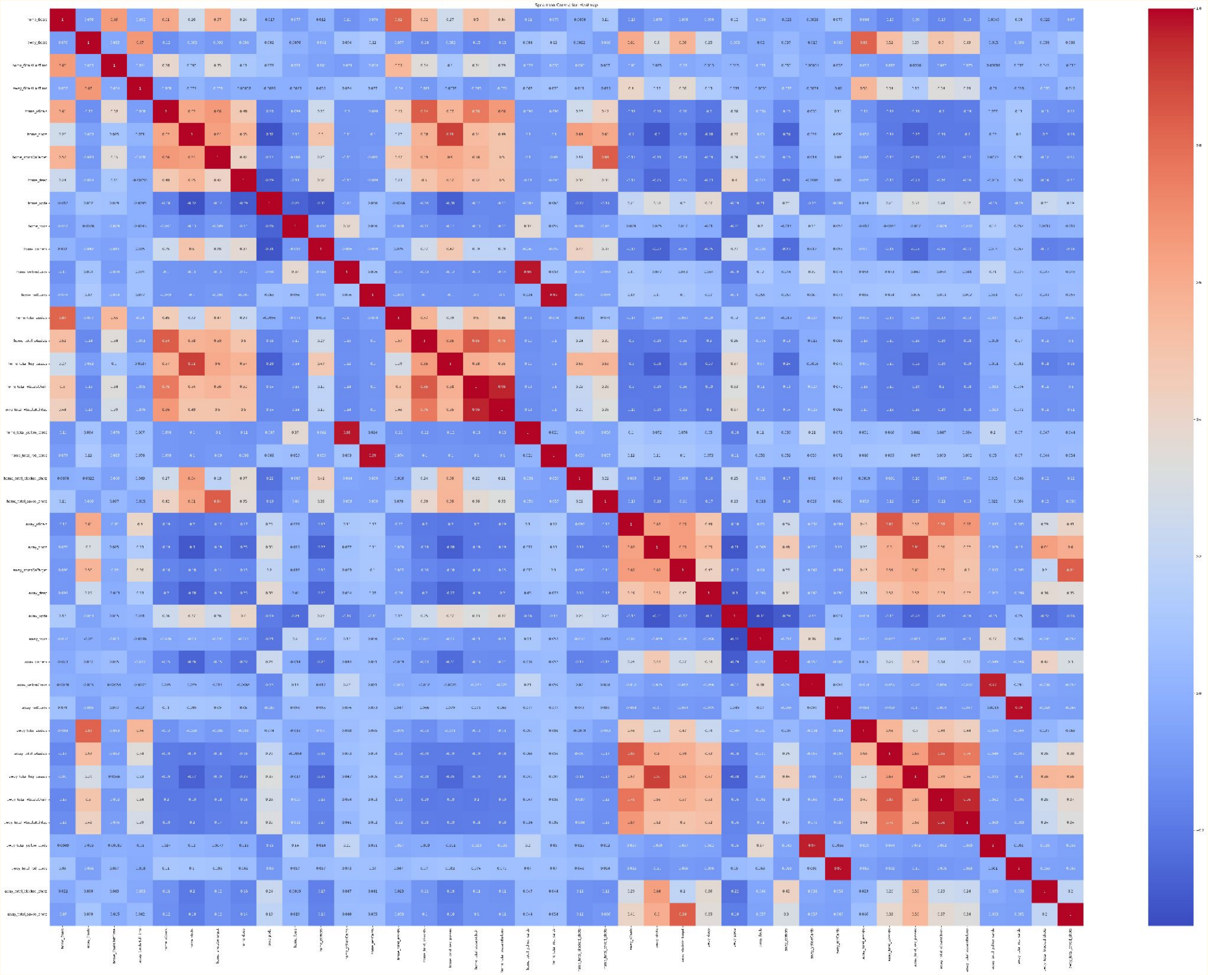
	playerID	teamID	playerName	teamName
0	560	89	Sergio Romero	Manchester United
1	557	89	Matteo Darmian	Manchester United
2	548	89	Daley Blind	Manchester United
3	628	89	Chris Smalling	Manchester United
4	1006	89	Luke Shaw	Manchester United
...	...	...	...	...
10101	7396	176	Loic Bessile	Bordeaux
10102	9566	175	Yanis Lhéry	Saint-Etienne
10103	9565	175	Mathys Saban	Saint-Etienne
10104	9568	181	Charles Costes	Dijon
10105	9567	181	Erwan Belhadji	Dijon

Procedure

3. EDA-

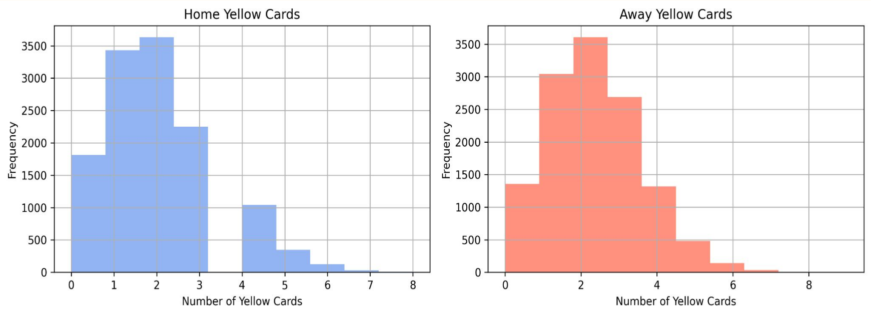
- Class balance: Nearly even across 3 outcomes
- Key variables:
  - Goals, Shots, xG(goals and assist), Corners, Cards
- Statistical tests:
  - T-tests for numeric vars
  - Chi-square for categoricals

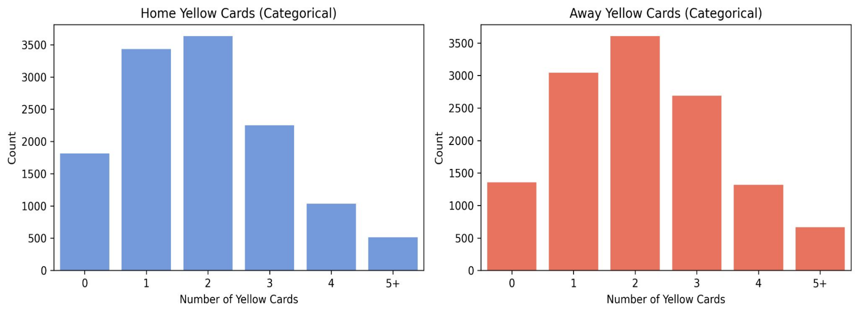
Correlations

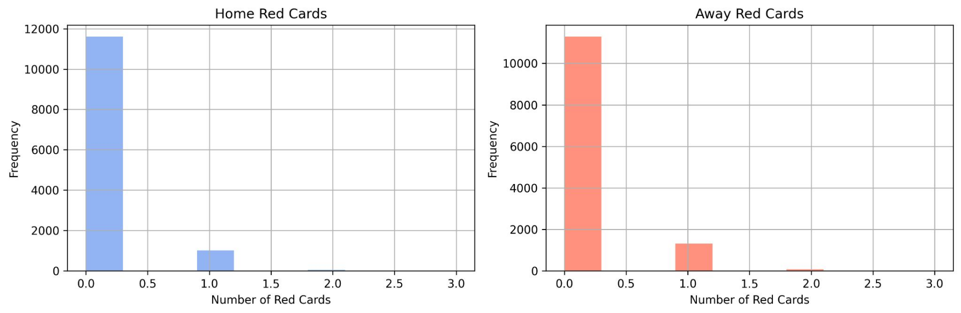


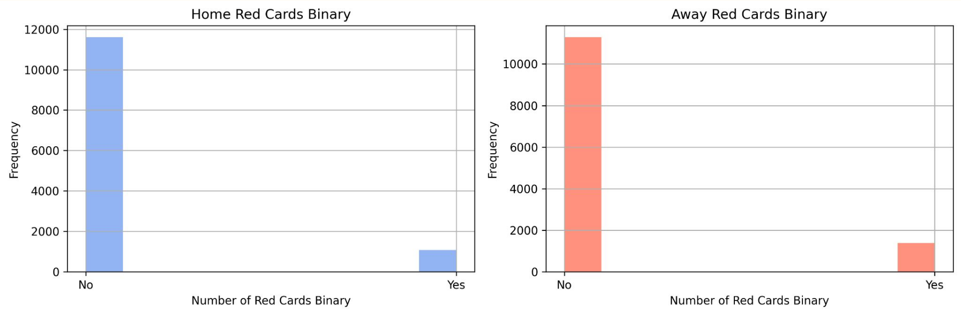
Procedure

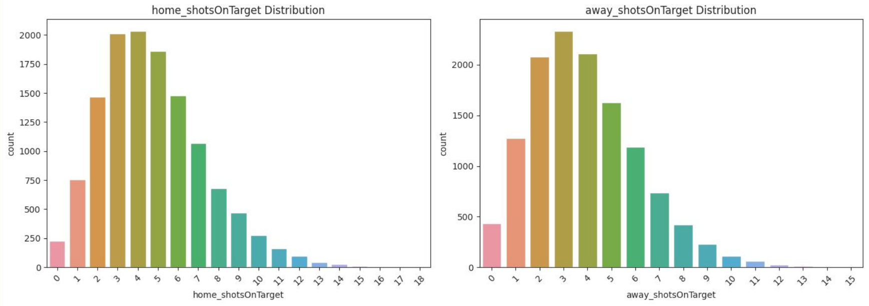
4. Creating categorical features to handle outliers with low cardinality

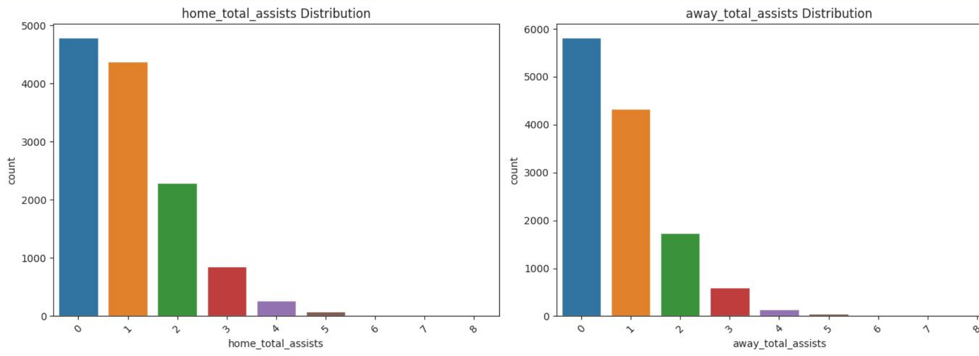


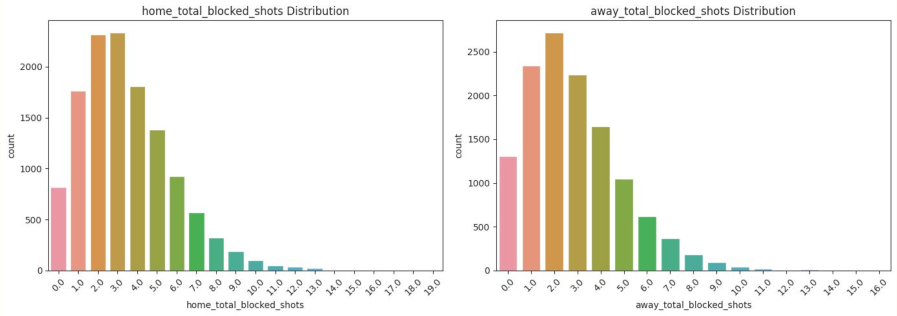


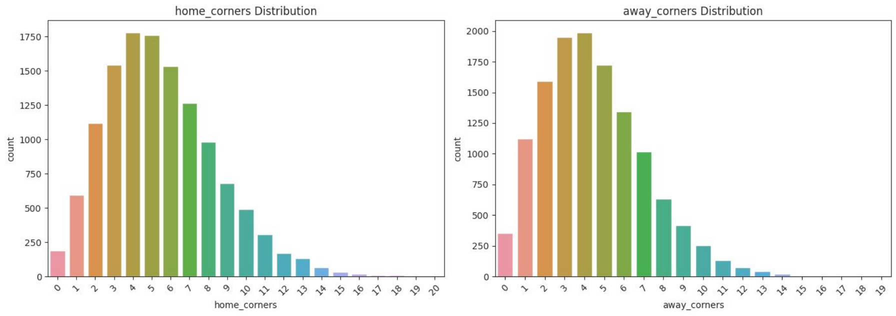


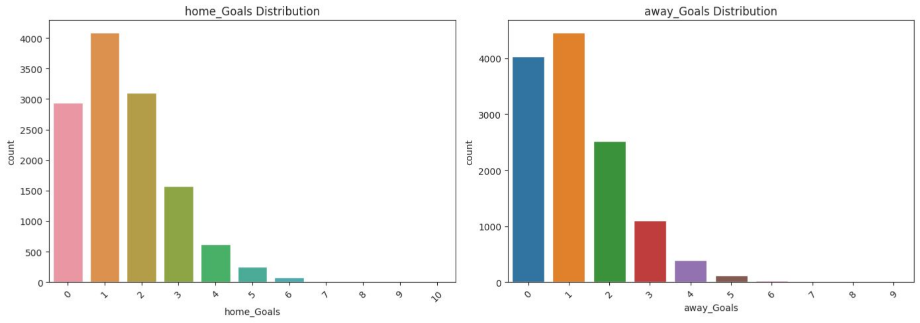


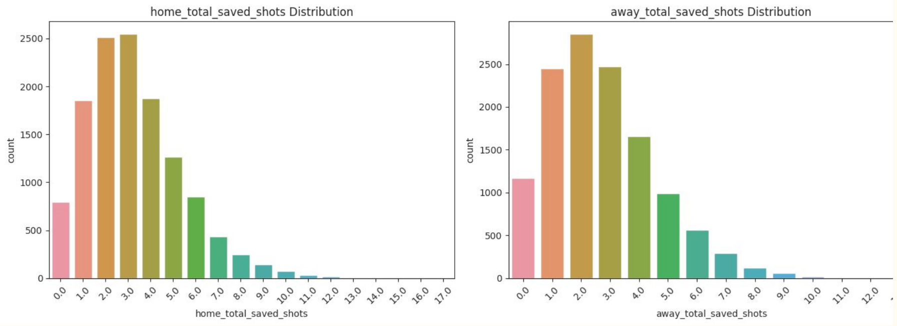

home_shotsOnTarget Distribution
count
0
250
500
750
1000
1250
1500
1750
2000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
home_shotsOnTarget
away_shotsOnTarget Distribution
count
0
500
1000
1500
2000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
away_shotsOnTarget


home_total_assists Distribution
count
0
1000
2000
3000
4000
5000
0 1 2 3 4 5 6 7 8
home_total_assists
away_total_assists Distribution
count
0
1000
2000
3000
4000
5000
6000
0 1 2 3 4 5 6 7 8
away_total_assists


home_total_blocked_shots Distribution
count
0
500
1000
1500
2000
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0 13.0 14.0 15.0 16.0 17.0 18.0 19.0
home_total_blocked_shots
away_total_blocked_shots Distribution
count
0
500
1000
1500
2000
2500
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0 13.0 14.0 15.0 16.0
away_total_blocked_shots

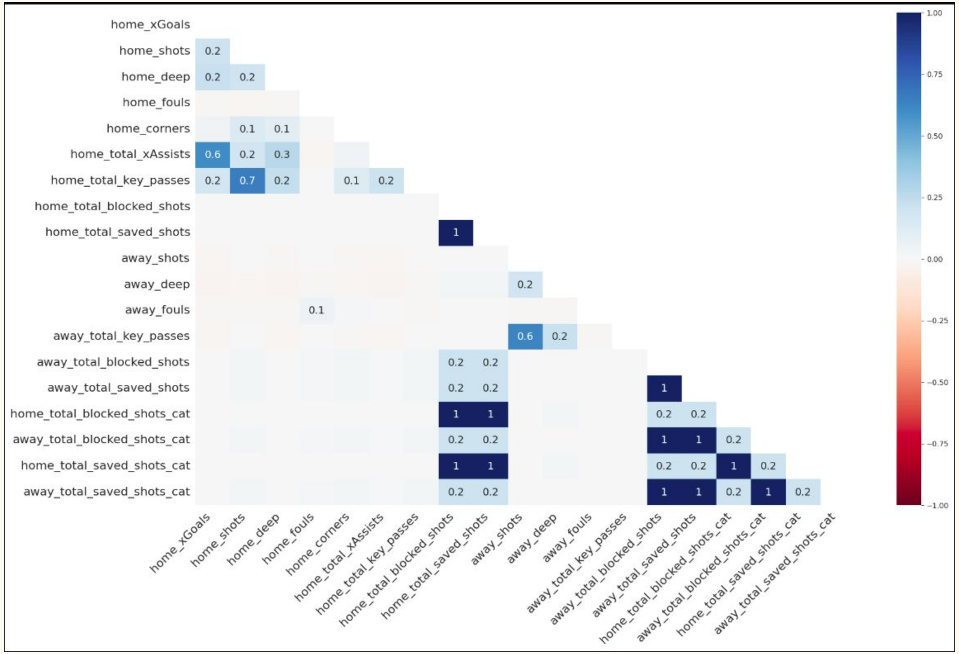

home_corners Distribution
count
0
250
500
750
1000
1250
1500
1750
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
home_corners
away_corners Distribution
count
0
250
500
750
1000
1250
1500
1750
2000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
away_corners


home_Goals Distribution
count
0
500
1000
1500
2000
2500
3000
3500
4000
0 1 2 3 4 5 6 7 8 9 10
home_Goals
away_Goals Distribution
count
0
1000
2000
3000
4000
0 1 2 3 4 5 6 7 8 9
away_Goals


home_total_saved_shots Distribution
count
0
500
1000
1500
2000
2500
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0 13.0 14.0 15.0 16.0 17.0
home_total_saved_shots
away_total_saved_shots Distribution
count
0
500
1000
1500
2000
2500
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0
away_total_saved_shots

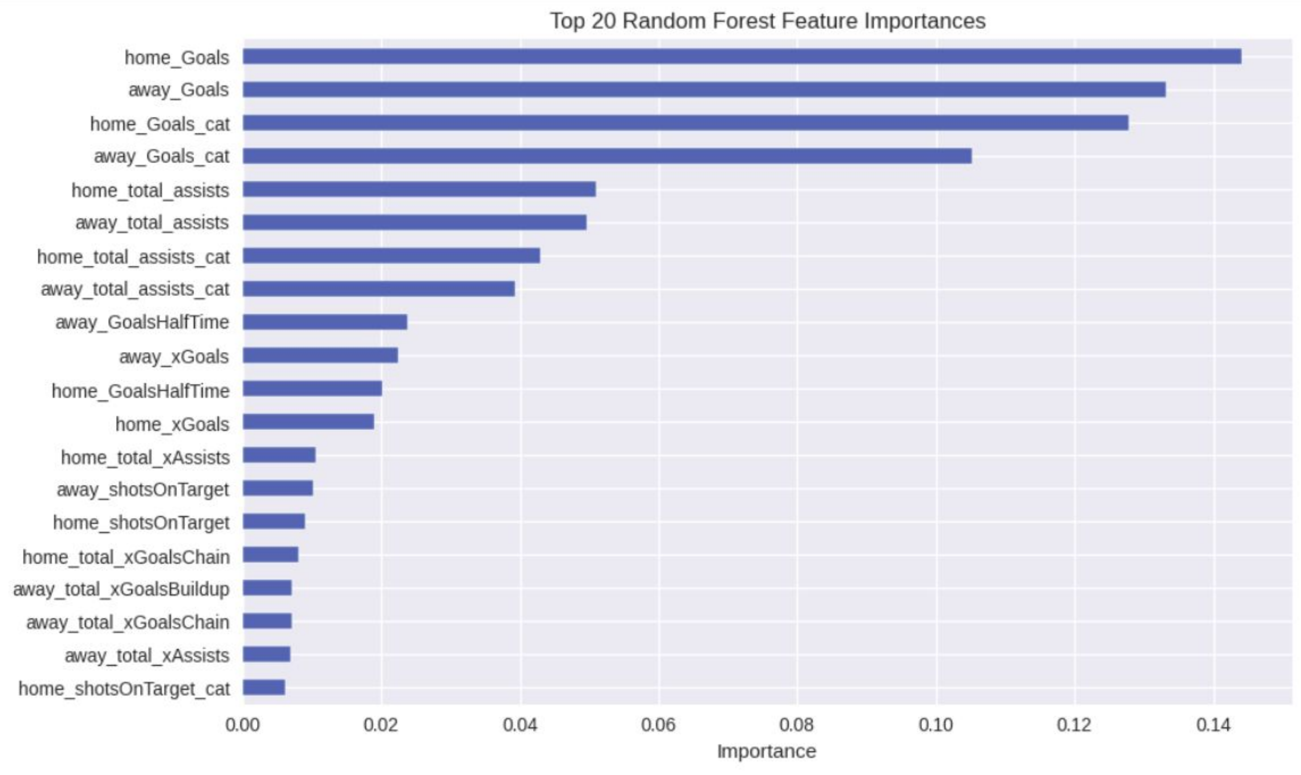
Procedure

5. Run MICE to fill missing values due to outliers removal



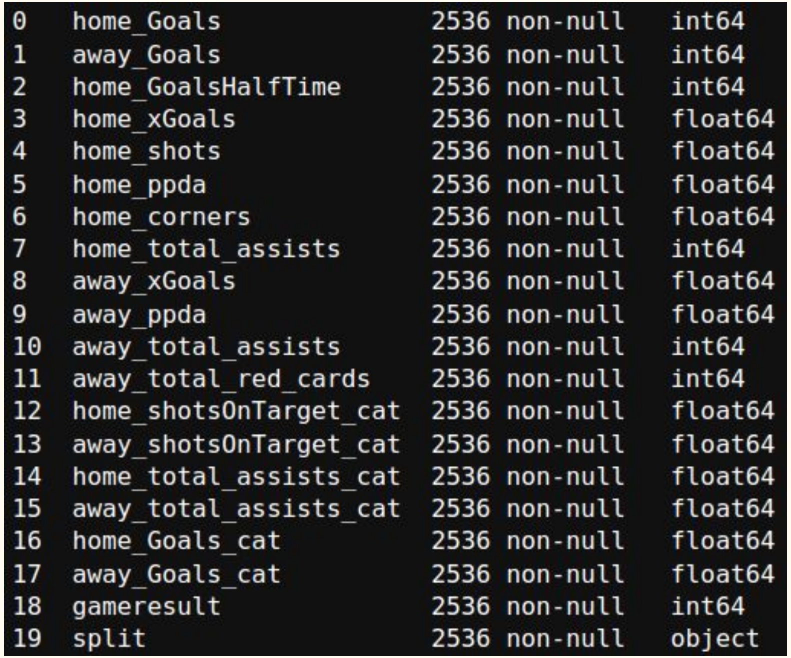
6. Feature engineering

- Created 30+ features:
  - Rolling averages
  - Goal efficiency ratios
  - Disciplinary scores
- Plotted distributions of engineered features



6. Feature Selection

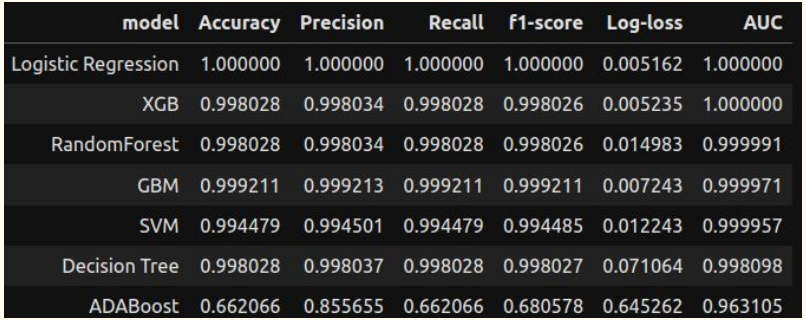
- Univariate tests (t-test, chi-squared)
- Multivariate: Lasso, Random Forest, Gradient Boosting
- Kept top 18 most important features



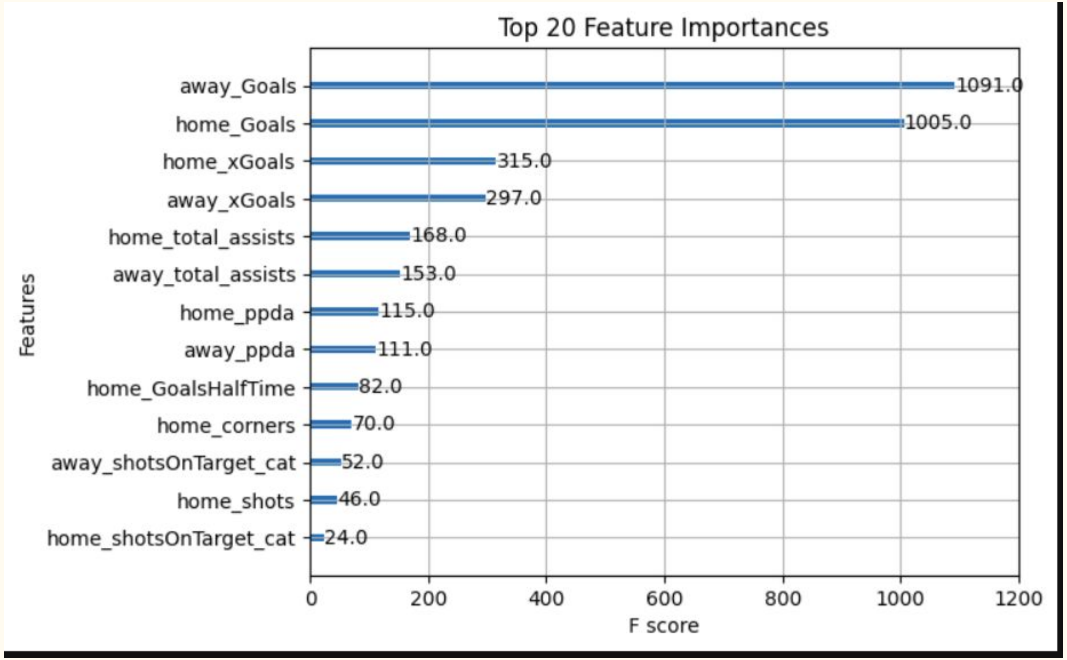
```
0   home_Goals              2536 non-null   int64
1   away_Goals              2536 non-null   int64
2   home_GoalsHalfTime      2536 non-null   int64
3   home_xGoals             2536 non-null   float64
4   home_shots              2536 non-null   float64
5   home_ppda               2536 non-null   float64
6   home_corners            2536 non-null   float64
7   home_total_assists      2536 non-null   int64
8   away_xGoals             2536 non-null   float64
9   away_ppda               2536 non-null   float64
10  away_total_assists      2536 non-null   int64
11  away_total_red_cards    2536 non-null   int64
12  home_shotsOnTarget_cat  2536 non-null   float64
13  away_shotsOnTarget_cat  2536 non-null   float64
14  home_total_assists_cat  2536 non-null   float64
15  away_total_assists_cat  2536 non-null   float64
16  home_Goals_cat          2536 non-null   float64
17  away_Goals_cat          2536 non-null   float64
18  gameresult              2536 non-null   int64
19  split                   2536 non-null   object
```

7. Modeling and predictions

Model: `XGBClassifier`



model	Accuracy	Precision	Recall	f1-score	Log-loss	AUC
Logistic Regression	1.000000	1.000000	1.000000	1.000000	0.005162	1.000000
XGB	0.998028	0.998034	0.998028	0.998026	0.005235	1.000000
RandomForest	0.998028	0.998034	0.998028	0.998026	0.014983	0.999991
GBM	0.999211	0.999213	0.999211	0.999211	0.007243	0.999971
SVM	0.994479	0.994501	0.994479	0.994485	0.012243	0.999957
Decision Tree	0.998028	0.998037	0.998028	0.998027	0.071064	0.998098
ADABoost	0.662066	0.855655	0.662066	0.680578	0.645262	0.963105



The project is not complete - definitely there is still work to do

Thanks